International Journal of Law in Context (2024), **20**, 568–580
doi:10.1017/S1744552324000156

ARTICLE

A comparative study on public interest considerations in data scraping dispute

Lanfang Fei

Law School, Jinan University, Guangzhou, Guangdong, China
Email: fei27@hotmail.com

(Received 25 June 2023; revised 29 February 2024; accepted 16 May 2024; first published online 06 September 2024)

Abstract

New technologies and the new business practices that they bring often raise difficult questions about the application of the law. This often stems from the difficulty of clarifying the impact of new technologies on the interests of different groups in society and, in particular, the difficulty of measuring the public interest brought about by new technologies. In practice, in disputes arising from new technologies, the users of the new technologies often justify their actions in terms of the public interest. In this paper, we compare data collection cases in the US, EU and China and extract the types of public interests discussed in typical cases in different countries, including (1) data-related property rights protection and commercial principles to prevent free-riding; (2) privacy, personal data protection and data security; (3) competition and innovation interests related to the free flow of data; and (4) freedom of expression. The comparison shows that an appropriate focus on the public interest in data flow has led Chinese and US courts to rule in favour of scrapers in a few recent cases, in contrast to the judicial attitude of EU courts, whicht value privacy. The author applauds the attitude of the courts in the US and China and argues that free competition and innovative interests based on data flows are public interests that should be prioritized in data scraping cases.

Keywords: data scraping; law; public interest; innovation

1 Introduction

In 1993, Matthew Gray, a PhD student at MIT, wrote the first web crawler program (Najork, 2009). Since then, scraping technology covering various technical terms, such as 'web scraping', 'screen scraping', 'data crawl and spider' and 'web bot', quickly became one of the most important and popular technical tools in the Internet and data industry (Udapure *et al.* 2014). The Imperva Threat Research (ITR) team reported that approximately 27.7 per cent of online traffic in 2022 came from automated bots (ITR 2022). Data scraping is the foundation of some important business models, including search engines, intellectual property checks, website or account aggregation, price comparisons and target advertisements (Dreyer and Stockton 2013). Any data, including user accounts or other personal information, product transaction records, customer reviews, blog posts, photos, videos and other web content, could be the target of a scraper (Grover *et al.* 2018). The relationship between data holders and scrapers in various business scenarios is complex. While some scraping services are beneficial to both parties, other scraping services may be detrimental to the interests of the data holder. This leads to disputes between data holders and scrapers (Krotov and Johnson 2022).

In 2022, the Ninth Circuit Court of Appeals of the US held that data scraping public websites is not unlawful, in *hiQ Labs, Inc. v. LinkedIn Corp.* (hereinafter referred to as the 'LinkedIn case'),

© The Author(s), 2024. Published by Cambridge University Press

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

which is viewed as the most emphatic pro-scraping ruling in the technology law area. Among other things, the Ninth Circuit found that giving 'companies like LinkedIn free rein to decide, on any basis, which can collect and use' publicly available data would risk 'possible creation of information monopolies that would disserve the public interest'.[1] Despite the concern of the public interest, the Court went so far as to enumerate a laundry list of claims that an aggrieved party might still assert against a scraper, and LinkedIn as the data holder finally obtained a favourable ruling and actually won a subsequent breach of contract lawsuit,[2] which made the judicial attitude even less clear. Accordingly, the question of what public interests are involved in data scraping and how the courts should consider these public interests remains unanswered.

This article will explore these questions by examining the categories of public interest involved in scraping technologies by looking at relevant cases in different jurisdictions, particularly the US, the EU and China, and on this basis, propose an approach to evaluating public interest in such kinds of disputes. This study makes several necessary and important contributions. First, recent legislative and judicial practice in the field of data scraping around the world shows some critical developments. The US Supreme Court narrowed the scope of the Computer Fraud and Abuse Act (CFAA)[3] to computer hacking in *Van Buren v. United States*, which seems to be favourable to third-party data users, including scrapers.[4] The European Commission published a proposal for a Data Act[5] that aims at enhancing data access and use within the EU in 2022. In China, the State Administration for Market Regulation (SAMR) issued a draft revision to the Anti-Unfair Competition Law (AUCL)[6] for public comments at the end of 2022, refining rules to regulate the improper acquisition or use of commercial data. Second, there are still some important debates in the academic community about the legality of data scraping. Scholars who advocate a lax attitude towards web scraping argue that scraping technology can help ensure the openness of information and prevent monopolies by large data platforms (Jamie 2018). They see data scraping as a technological tool that can contribute to competition by reducing information barriers for startups, correcting algorithms, providing new price comparisons or data-driven business services or promoting data-driven cultural and creative offerings (Andrew 2022). In contrast, commentators who advocate a more egalitarian approach to web scraping argue that this practice leads to breach of contract, free-riding, and invasion of users' privacy (Gold and Latonero 2017). Some even argue that courts should apply anti-trust law to prescribe public data scraping prohibitions imposed on dominant Internet companies (Ioannis 2019). In general, the topic involves several core questions about the technical rules and policy direction of the Internet, including who has the right to decide how public data are collected and used and how privacy and competitive interests can be reconciled.

The findings of this review are threefold. First, there is disagreement and controversy in the handling of data scraping cases by different courts in the US, the EU and China. These disagreements and controversies stem, on the surface, from the interpretation of the terms of specific legal provisions but, in essence, from the different weighing of private and public interests by the courts in the context of unclear data ownership. Second, the US and Chinese courts have developed common law approaches to data scraping cases under the legal frameworks of computer trespass law and unfair competition tort, respectively. In contrast to the approaches of the US and China, the EU's strong data protection policies have led to a reduction in such disputes and a lack of case-by-case exploration of the issue in court practice. Third, there have been a few pro-scraper decisions made by courts in both China and the US, allowing scraping without

[1] *hiQ Labs, Inc. v. LinkedIn Corp*., No. 17-16783 (9th Cir. 2022).

[2] *hiQ Labs, Inc. v. LinkedIn Corp*., No. 17-3301 (N.D. Cal. Nov. 4, 2022)

[3] The Computer Fraud and Abuse Act, 18 U.S.C. § 1030.

[4] *Van Buren v. United States*, 141 S.Ct. 1648 (2021)

[5] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on harmonised rules on fair access to and use of data (Data Act)COM/2022/68 final

[6] Draft Revision to the Anti-Unfair Competition, issued by the SAMR on November 22, 2022.

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

substantial injury to the data controllers by referring to public interests in data sharing. The author applauds this tendency. The case review reveals that the conduct related to data protection is not harmful per se to competition but instead may contribute to innovation and be beneficial to consumers. With this in mind, this paper proposes an approach for the appropriate consideration of the public interest of data flow in multiple legal contexts.

The rest of the article is divided as follows: Part I provides the technical and commercial landscape of data scraping and outlines the conflict of private and public interest arising from the technology. Part II introduces the relevant laws and data scraping cases in the US, the EU and China. Part III compares the differences in the ways that the US, the EU and Chinese courts weigh and consider public interest in dealing with data scraping cases, discussing how the courts in the three jurisdictions have reached different conclusions. Part IV discusses an appropriate framework for considering the public interest in scraping cases. Part V is the conclusion.

2 Conflict of interest arising from data scraping

There are various categories of automated software that can be used to exploit data scraping by extracting data from sites, platforms or mobile applications. Data scraping can also be performed manually by Internet users or done in an automated manner using a specialized program known as a web spider or bot. In general, the job of data scraping is rather similar to copying data from websites or other sources into appropriate data structures, such as spreadsheets and relational databases (Rahman and Tomar 2021). Regardless of the different technical routes, data scraping is, by its very nature, a method of collecting data and information. Just as the judge in *Sandvig v. Sessions* recognized, scraping 'is merely a technological advance that makes information collection easier; it is not meaningfully different from using a tape recorder instead of taking written notes'.[7] The scraped data can be used for positive purposes, such as producing personalized products or providing customized service, predicting market changes, optimizing producing structure, formulating price plans, arranging competition strategies and doing research. The scraped data can also be used for negative purposes, including establishing fake websites, committing intellectual property or trade secret theft and causing difficulties in accessing websites by traffic attack (McRory 2021).

Data scraping involves different stakeholders in different business scenarios, but the direct stakeholders involved in this technology are inevitably data controllers and scrapers (Hirschey 2014). A data controller might be the owner of a website, digital platform, mobile application, etc., who provides technical support for data aggregation. A data controller could also be a data contributor or data user. This party has a corresponding use and revenue interest based on the data under its control. Another party is data scrapers, who wish to extract website or platform data by bot, spider or other automatic programs for different purposes. While some scraping services are beneficial to both parties, other scraping services may be detrimental to the interests of the data controller. That is, the data product that the scrapers create with data scraping can directly or indirectly compete with the website controller's business, resulting in an unfair distribution of value from the data controller (Wierzel 2001). Scraping can also slow down a website owner's servers and increase webpage load times, negatively impacting user experience and the website's revenue stream. In addition to direct stakeholders, data scraping may concern other parties, including the registered users of the website or platform, consumers and any other relevant interested entities. These people may be contributors and creators of data on webpages or platforms. Data scraping may, on the one hand, negatively impact their privacy, the security of their personal information and the ease of access to webpages and, on the other hand, positively impact the benefits of access to competitive or innovative products and services, as well as the fair use of data and free speech. In general, the data scraping issue is complicated because it involves a

[7] *Sandvig v. Sessions*, 315 F. Supp. 3d at 1 (D.D.C. 2018).

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

set of conflicts between traditional business ethics prohibiting free-riding and the competing demands for open access to data, interspersed with nuanced trade-offs among privacy, data security and the monopoly power of platforms.

The conflicts of private and public interest in data arising from scraping conduct have led to a number of disputes all over the world. The disputes embed a value choice informed by the commercial interests of data controllers and scrapers. To date, no country has had a specific legal rule governing data scraping, nor has any country explicitly considered data scraping to be a per se prohibited conduct. Accordingly, data scraping may give rise to various and overlapping legal claims and legal liabilities (Kerr 2016). Scraping conduct can give rise to criminal liability if it has socially harmful consequences, including offences such as intrusion into computer systems, trade secrets and personal information (Tsang 2022). The relevant legal provisions exist in national criminal laws all over the world. In the civil realm, data scraping, depending on the specific scenario in which it is committed and the relevant facts, may inspire different claims, including intellectual property infringement, breach of contract, trespass infringement, unfair competition and monopoly. The diversity of the data scraping and relevant business models brought about by them, together with overlapping legal sources, often raise difficult questions about the application of the law. This often stems from the difficulty of clarifying the impact of scraping conduct on the interests of different groups in society, particularly the difficulty of measuring the public interest brought about by scraping. In practice, in disputes arising from new technologies, both data controllers and scrapers often try to justify their actions on the grounds of public interest. The data involved in the data scraping cases we discuss include personal or non-personal data or even both. Thus, the types of public interest defenses raised by parties in different data scraping cases may vary accordingly. When personal information is involved, the primary public interests involved in scraping are privacy and personal information protection. If non-personal information is involved, the primary public interest at stake is the innovation and consumer welfare resulting from data sharing. The following section introduces and compares what kinds of public interests have been raised in scraping disputes and how courts consider the public interest factor in the US, the EU and China.

3 Public interest considerations in scraping disputes

The issue of public interest in the Internet, data and relevant technologies is a broad topic (Elkin-Koren 2020). Much of the literature has analysed the various impacts of the Internet and digital technologies on public benefits at a macro level, including competition and innovation, total social welfare, privacy protection, freedom of expression, online democracy, environmental protection, etc. (Ingrams 2019). This paper will examine the public policy that presents itself in specific disputes from a micro-case perspective. For this purpose, the paper defines public interest as the impact of digital capture on the interests of third parties rather than on those of the direct stakeholders, with reference to what US courts have found, specifically that '[t]he public interest inquiry primarily addresses impact on non-parties rather than parties', as set out by a US court.[8]

As mentioned above, data scraping may give rise to various and overlapping legal claims and legal liabilities. Despite the complex legal sources, the most popular laws to which parties resort in scraping disputes may vary due to national legal traditions, regulatory framework and cost-effect constraints in different jurisdictions. However, different laws leave room for the public interest in the constituent elements, typically the fair use doctrine in copyright law, the only difference being the amount of room. National courts have had to analyse disputes involving scraping in light of various legal sources and have made divergent decisions (Goldfein and Keyte 2017).

[8] *Bernhardt v. Los Angeles County*, 339 F.3d 920, 931 (9th Cir. 2003)

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

3.1 The US courts: Limiting the use of CFAA for scraping conduct

For a long time, the US courts supported the interest of data holders by applying the Computer Fraud and Abuse Act (CFAA), an anti-hacking law that sets a lower threshold for web scraping than other substantive causes of action. Specifically, the CFAA establishes civil and criminal liabilities for anyone who 'intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains ... information from any protected computer'. Other common law or statutory claims, such as trespass to chattels, tort interference with contract or unfair competition, were often subordinated to the CFAA in litigation (Sobel 2020).

The first data scraping case heard by a US court was *eBay v. Bidder's Edge* in 2000, in which eBay asserted a trespass to chattels claim and a claim under the CFAA. A federal district court issued a favourable injunction for eBay on the grounds that web aggregation as a negative interference would substantially impair eBay's computer system. Andrew Sellars collected over 60 data scraping rulings under CFAA and analysed the changing attitudes of the courts (Sellars 2018). In the early days, the courts largely interpreted CFAA's requirement of 'without authorization' to mean that permission was not granted at the time of access. On this understanding, it is sufficient for the website owner to prove that it does not wish to authorise the scraper to capture the data, either by posting a unilateral statement on the website or expressing restrictions on scraping in the website contract. The courts' attitudes began to split in 2010. Some courts began to reject the argument that a breach of the terms of use of a website constitutes a breach of the CFAA and started to emphasise that only scraping that breaks technical limitations can constitute 'without authorization'. In addition, courts have begun to distinguish between public and nonpublic data and to emphasise the distinction between 'use' and 'access'. For instance, in the case *Facebook, Inc. v. Power Ventures, Inc.*, the Ninth Circuit noted that the CFAA does not prohibit data scraping of public information on the sole basis of terms-of-service violations. However, as most courts still consider cease and desist letters sufficient to prove 'without authorization', it is still difficult for scrapers to survive in CFAA cases. In several cases, the courts confirmed that web scraping constituted 'unauthorized access' after an express revocation of permission violated the CFAA.[9]

The *LinkedIn case* marked a turning point, as the Ninth Circuit finally rendered a favourable decision for the scraper. In the *LinkedIn* case, both parties put forward several public interest arguments to defend themselves, and the courts responded to these arguments. These public interests included the following: (1) Privacy protection. Regarding the balance of the equities, the Ninth Circuit asserted that LinkedIn's interest rooted in the users' privacy in preventing hiQ from scraping its data was not sufficient to outweigh hiQ's interest in continuing to conduct its business by relying on access to and use of data from LinkedIn's public website. (2) Preventing free-riding. The Court also did not accept the argument that LinkedIn had a legitimate interest in preventing 'free riders' from using data on its platform. The Court said that LinkedIn 'has no protected property interest in the data contributed by its users' and that LinkedIn could satisfy its 'free rider' concern by eliminating public access. The Court further denied the 'legitimate business purpose' argument under the tortuous interference raised by LinkedIn, saying that LinkedIn's core business model did not require prohibiting HiQ from using its platform's publicly available data. (3) Data flow and relevant competition interest. Each side asserted that its own position would benefit the public interest by maximizing the free flow of information on the Internet. While the Ninth Circuit admitted that there were significant public interests on both sides, it asserted that public interest favoured granting the preliminary injunction, as it would not prevent LinkedIn from continuing to engage in 'technical self-help' to defend itself against malicious behaviour. (4) Free speech. During oral arguments, hiQ argued that social media sites such as LinkedIn are the modern equivalent of the town square and that allowing LinkedIn to choose who can access the

[9]For example, EFCultural Travel BV v. ZeferCorp., 318 F.3d 58, 62 (1st Cir. 2003), Southwest Airlines Co. v. Farechase,Inc., 318 F. Supp.2d 435, 439–440 (N.D. Tex. Mar. 19, 2004), CollegeSource,Inc. v. AcademyOne,Inc., 597 Fed.Appx.116, 130 (3d Cir. 2015), EarthCam,Inc. v. OxBlueCorp., 703 Fed.Appx.803, 808 (11th Cir. 2017).

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

site is a violation of the First Amendment and will have grave constitutional consequences. LinkedIn countered that its platform is not a public square but, rather, is similar to a public library and that borrowing books requires compliance with the rules of the platform. The Court did not clarify the nature of the platform providing public data but instead emphasized that allowing the data holder to determine how to share the data on publicized websites might allow website owners to block access for 'discriminatory, anticompetitive, or other improper reasons.'

In general in this case, the Ninth Circuit seemed to shift from a literal interpretation of the CFAA to a legislative purpose interpretation, explaining through a review of congressional legislative intent that the CFAA should not be used against publicly available data on the Internet. This approach would significantly limit the scope of the CFAA's application to civil disputes relating to data scraping.

3.2 The EU courts: Imposing openness obligation by competition law

Before the draft of the General Data Protection Regulation (GDPR), courts in the EU, including other European countries, had different views on data scraping, considering the public interest in the context of intellectual property law and unfair competition law. These public interests include the following: (1) Temporary copies and fair use of copyrighted data. For instance, in a case concerning the use of crawling and scraping of news websites by the operator of a paid news aggregator service that allowed its users to read extracts of the crawled articles, the UK Supreme Court asserted that the scraping conduct as the act of temporary reproduction would fall within the fair use exception for temporary copying, as it was in the service of a lawful act and was temporary.[10] That ruling was finally affirmed by the Court of Justice of the European Union (CJEU).[11] By referring to the preamble to Directive 2001/29, which sets out '[A] fair balance of rights and interests between the different categories of rightholders, as well as between the different categories of rightholders and users of protected subject-matter, must be safeguarded ... ',[12] the CJEU agreed that the on-screen copies and the cached copies do not unreasonably prejudice the legitimate interests of those rights holders. In short, products or works that use scraped data for non-commercial research, educational or other transformative purposes are likely to be protected by the fair use doctrine under copyright law. But fair use is determined on a case-by-case basis, and a particular scraping must meet the elements of fair use in order to establish an exception. (2) Increasing competition and consumer welfare. German courts had analysed the conflict of interest in data scraping in a more substantial way under unfair competition law.[13] In a 2010 case, the German Federal Supreme Court (BGH) in the Automobil-Onlinebörse case held that a software provider that can compare car advertisements on different websites does not use the data itself and therefore cannot be a direct infringer.[14] The German court also rejected the argument that the operation of a meta search engine with an integrated booking system would constitute an act of unfair competition only because the source website did not allow this practice in its general terms and conditions in 2014. The Court came to this conclusion after balancing the interests of all concerned parties, including consumers at large, emphasizing the positive effects that price transparency has on competition.[15]

In contrast, the CJEU, in cases without copyrights, may have a tough attitude towards scrapers. The CJEU ruled negatively against a meta search engine operator who searched for and collected car sales advertisements from third-party websites. The Court argued that the search and

[10]*NLA v Meltwater*, Case C-360/13.

[11]*Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd and Others*, Case C-360/13.

[12]Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

[13]Act against Unfair Competition (UWG), effective on 28 May 2022.

[14]*Urteil des V. Zivilsenats vom* 11.6.2021 - V ZR 234/19

[15]Flugvermittlung im Internet, Case No. I ZR 224/12.

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

aggregation conduct constitutes 're-utilisation' and infringes on the *sui generis* right of the web owner. The Court argued that the defendant leads risk that the database would lose income and that the conduct was equal to the manufacture of a parasitical competing product.[16] Moreover, it also ruled negatively against scraping in *PR Aviation v. Ryanair*; the Court decided that the company scraping Ryanair's database of flight times and prices had not infringed copyright because the data involved were lacking requisite creative input. The CJEU made it clear, however, that it is possible for a website owner to restrict the reuse of the mined data through the terms and conditions in its contracts.[17] This view essentially supports website owners' ability to counter data scraping by setting unilateral contractual terms, regardless of whether the data group has database rights. This view is similar to the earlier negative view of data scraping held by the US courts.

Since 2018, the rules concerning the processing of personal data in EU Member States have been unified by the GDPR. The GDRP sets out rules governing the processing of personal data and defines the legitimate interest of data controllers and processors to a certain extent. In the framework of the GDRP, data scrapers are required to fulfil a relatively heavy burden, while the right of data probability might imply that a scraper could transfer data with the user's consent. Law enforcement agencies often rely on the GDRP to impose liability directly for scraping without user consent. Strict personal data protection laws have somewhat limited litigation between data controllers and scrapers, which may explain why civil disputes over data scraping have been less frequently reported in Europe in recent years.

Notably, Europe appears to be attempting to protect the public interest in data sharing by creating openness obligations for large digital platforms such as so-called gatekeepers. On 1 November 2022, the EU Digital Marketplace Act (DMA), which includes far-reaching interoperability obligations for messenger services, came into force.[18] The EU also proposed a Data Act that envisages interoperability provisions, especially for cloud services.[19] However, it remains to be seen how and to what extent the gatekeeper should achieve interoperability and how it would impact data scraping. Another related legal act, the EU Data Governance Act, also came into force in 2023, setting out the conditions for the reuse of data held by public sector bodies. Overall, the latest legislative developments in the EU demonstrate a public interest bias towards data sharing and innovation. Moreover, these developments offer greater possibilities and scope for the legitimization of data scraping from different perspectives. Data scrapers may be able to justify their actions on the basis of platforms' interactivity obligations, contractual fairness clauses and public data sharing clauses. However, the exact conditions under which these clauses apply are unclear, and the way in which they affect data scraping in practice will have to be clarified in specific cases.

3.3 The Chinese courts: Considering public interest under Anti-Unfair Competition Law

China deploys unfair competition law to address data scraping cases. China has two principal laws that address different aspects of competition. The Anti-Unfair Competition Law (AUCL) covers a range of unfair competition practices, with a focus on the fairness of business transactions and business ethics.[20] The main laws relevant to data scraping in China are Articles 2 and 12 of the AUCL. The second paragraph of Article 12 of the AUCL prohibits operators from using techniques to prevent or disrupt the normal operations of network products or services that are

[16]*Innoweb BV v.Wegener ICT Media BV*, Case C-202/12.

[17]*PR Aviation v. Ryanair*, Case C-30/14.

[18]Digital Markets Act (DMA), published in the Official Journal of the European Union on 12 October 2022 and entered into force on 1 November 2022.

[19]Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access and use of data (Data Act) COM(2022) 68 final.

[20]Anti-Unfair Competition Law, promulgated by the Standing Comm. Nat'l People's Cong., revised in 2017,2018 and 2019, effective April.1, 2019) (China)

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

lawfully provided by other operators' network products or services provided by the normal operation of the behaviour.[21] From the context of the article, the conduct subject to the regulation of the article needs to meet the requirement that other operators have legitimate network products and services, using technical means, and cause obstructing and disrupting effects. Since it is difficult for data controllers to prove the obstructing and disrupting effects, data controllers are more likely to resort to Article 2, otherwise known as the general clause, which provides that business operators shall follow the principles of voluntarism, equality, fairness and good faith and abide by the law and business ethics. Article 2 further defines unfair competition activities as those that disrupt the order of market competition, thus harming the legitimate rights and interests of other operators or consumers. The formulation of the provision and the elements needed to cover the 'legitimate rights and interests' is extremely broad and requires specific behaviour in violation of the law and business ethics, disrupting the order of competition and harming other operators and consumers.

The earliest case in China relating to data copying and crawling was *Aibang v. Hantao* in 2011, in which the defendant used vertical search engine technology to copy the shop descriptions and user reviews of the plaintiff's website. The Court held that the plaintiff's conduct disrupted the economic order in the online environment and harmed market competition, which constituted unfair competition prohibited by Article 2 of the AUCL. With the rapid development of China's Internet industry, disputes related to data crawling have increased rapidly. Among these cases, most plaintiffs are large Internet platform enterprises that hold massive data resources. Like their US counterparts, Chinese courts provided an early treatment of data scraping that was largely in favour of data controllers, requiring the scrapers to obtain authorization from the data controller. A typical example of this attitude is the landmark case *Weibo vs. Maimai*, in which the Court established a triple authorization framework, only by which a complete chain of rights for grabbing data through Open API can be achieved: first, the platform shall be authorized by users to acquire their personal information; second, a third-party developer shall be authorized by the platform; finally, the third-party developer shall be authorized by the users.[22] The triple authorization doctrine imposed extremely strict requirements on scrapers. Under this doctrine, crawlers need be not only the explicit authorization of the website owner but also the explicit authorization of the website user. Both authorizations are necessary but not sufficient conditions, and neither is indispensable. This is almost impossible to achieve in practice. Under this rule, data crawling issues are examined from the perspective of permission without any distinction of the types of the data or the intention and effects of the scraper, leaving the data controller to decide the processing of the data in practice. In a number of subsequent scraping cases, courts have endorsed the judicial framework and found that scraping constitutes unfair competition on the grounds that the scraper was not authorized by the website owner and user.[23]

In recent years, however, several courts have gradually begun to consider more public interest implications when dealing with data scraping cases. The main public interest factors considered by Chinese courts in cases include the following: (1) Privacy and data protection. The right to privacy is, to some extent, the foundation of all data rights. In early cases, courts generally considered privacy and personal data protection–related rights to be important factors in determining the legitimacy of competitive conduct. Some courts identified that the breach of the privacy agreement between the data controller and the user is one of the harmful consequences of unfair competition. For example, in *Weimeng v. Fuyu*, the Court held that the data controller has the obligation to protect user data security or pay consideration according to the bilateral agreement, while the defendant's scraping and displaying of the data of the social platform involved would inevitably

[21]ACUL, art.12.

[22]*Maimai v. Weibo*, (2016) Jin 73 Civil Final No.588.

[23]For example, *Tencent v Yundian*, (2019) Yue 03 Civil TriaTrai No. 1913 ; *Tencent v Heju*, (2020) Guangdong 0104 Civil Judgment No. 46873.

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

negatively affect the performance of the agreement between the platform and its users.[24] In another case, *Tencent vs. Dianyun*, the Court held that the account and password of the user are the personal data of the game player, which should be controlled and used exclusively by the player.[25] (2) Free choice and interests of consumers. Few courts have compared the benefits to consumers and the losses to data controllers. In the *Qianjing v. Yichang* case, the Court considered the user's will and choice, the convenience to the consumer, and the limited loss of traffic to the data controller. With respect to the linked extranet account service, the Court found that it was the choice of the defendant's business users whether to link, that the technology itself was not improper and that the service could bring convenience to the users and actually increase the benefits to consumers.[26] (3) Competition and innovation. For instance, in *Tencent vs. Sishi*, the Court proposed an 'undistorted competition standard', which advocates that the definition of unfair competition must be prudent and neutral, without predicting the outcome of competition and without intervening in the process of competition. The Court pointed out that data scraping in general may discourage the creativity of online platforms and distort the competition mechanism of the big data market, while data scraping can also promote information disclosure. However, the Court did not make a detailed measurement of the benefits and disadvantages.[27]

4 Observations of international development

The above review of the US, the EU and Chinese jurisprudence on data scraping seems to indicate that the issue is still unclear (Riley 2018). The courts in different jurisdictions make decisions based on different legal claims and policy considerations (Luscombe *et al.* 2022). While competition law may point in one direction, trespass law points in another, and contract law points in yet another. The US courts' approach to data scraping has taken several turns over the 'with authorization' under CFAA. While the Ninth Circuit's discussions over injunctions in the *LinkedIn* case seemed friendly to scrapers, the eventual hiQ failure in that case and recent judicial attitude show that data can still restrict data crawling based on unilateral contractual terms. The interpretations of German courts based on unfair competition and of EU courts based on contractual terms also split on the legality on data scraping, and there is no further clarification on the convergence due to the intensive data protection under the GDRP. The Chinese courts appear to have partially abandoned the strict authorization requirement in favor of a broader approach of interest weighing, while the specific weighing of interests appears to be left entirely to the discretion of the judge, resulting in a number of contradictory decisions (Martens and Zhao 2021). The complexity of the problem is exacerbated by the diversity of legal relations and by the corresponding legal rights involved in different types of data (Mancosu and Vegetti 2020). In the face of this challenge, courts have often had to deploy general public interest considerations under existing legal rules to address the various conflicts arising from data competition.

The following table indicates the differences and similarities of the three jurisdictions from a public interest perspective (Table 1).

Even with this lack of clarity, we can still find some valuable information based on a comparison of these three jurisdictions. Among other things, one of the most notable findings is that courts in both the US and China are showing signs of relaxing judicial constraints on data scraping, with courts in both countries siding with data scrapers in a small number of cases. There are also similarities in the rationale deployed by the courts in those cases favourable to scrapers. First, these pro-scraping-friendly courts have commonly held that the data controller does not necessarily have a legitimate interest in the data under control. Second, both US and Chinese

[24] *Weimeng v. Fuyu Fanyou*, (2019) Jing 73 Civil Final No. 2799.
[25] *Tencent vs. Dianyun,*, (2020) Zhe 0192 Civil No. 1330.
[26] *Qianjin v. Yicheng*, (2019) Shanghai 73 Min Final 263.
[27] *Tencent vs. Sishi*, (2021) Zhe 8601 Civil Trial No. 309.

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

Table 1 Judicial attitude on public interest on data scraping disputes

	United States	EU	China
Major applicable law	CFAA	Copyright and *sui generis* law	AUCL
Types of public interest considered by the courts	A. Privacy protection B. Preventing free-riding C. Data flow and relevant competition interest D. Free speech	A. Temporary copies and fair use B. Increasing competition and consumer welfare	A. Privacy and data protection B. Free choice and interests of consumers C. Competition and innovation

courts have conducted a more detailed analysis of whether scraping constitutes illegal trespass or interference. Some courts no longer consider that scraping data controlled by other parties necessarily constitutes unauthorized intrusion and interference. Third, some courts require the data controller to prove that it suffered some actual, substantial harm to its business that was caused by the scraper activity rather than by traffic loss. In addition, courts in both the US and China have taken note of the public interest issues involved in data theft cases. In general, the US and China have developed common law paths for data scraping, i.e. approaches that consider the public interest in addition to data scraping conduct. Moreover, the courts in China and the US appear to be moving from different starting points to the same destination: to encourage data flow rather than private interest protection. In contrast to China and the US, EU courts clearly take a more severe approach to data scraping, although there are a small number of cases in German courts upholding data scraping under unfair competition law. Most of the laws applied by EU courts are similar to intellectual property laws or database laws, and the public interest is largely explored within the framework provided by these laws themselves. While there has been much legislative effort to promote data sharing by creating openness obligations for large digital platforms, this effort has not yet been reflected in actual disputes.

5 Some reflections on considering the public interests

The review of the data scraping cases shows that what appear to be private disputes among hucksters almost invariably touch public welfare (Brown 1948). There are a variety of public interests surrounding data capture disputes. The final questions, then, are how these different public interests are prioritized and how they should be considered in actual disputes.

First, among the various public interests concerning data scraping, we think that free competition and innovation based on data flow should be fundamental. By their nature, data constitute a kind of public good. Specifically, data are produced and created by both users and platform business operators. In the process of data production, users generate various data through registration, login, social interaction and the use of platform services, while the platform provides support for the presentation and expression of data (Krämer and Wohlfarth 2018). Business operators visualize data or other new products by arranging, sorting and processing so that potentially static data can flow and trade, generate value growth and finally realize economic benefits through consumption behaviour (Xiao 2020). Accordingly, data are a resource to generate incentives to create. The public interest in free competition has always been attacked by intellectual property protection or the exclusion of free-riding (Vezyridis and Timmons 2019). However, it should be noted that competitive markets do not operate because producers control the full social value of their output but because producers are allowed to receive a reasonable return. Allowing data appropriators to capture returns does not equate to allowing all the social value of the data under their control (Mayernik 2017). As long as the benefits cover the costs, uncompensated negative externalities do not prevent data controllers from making effective investments (Lemley 2005). The choice of applicable law in different countries is itself a testament

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

to the bias of interest. In general, computer trespass law, contract law, intellectual property law, unfair competition law and anti-monopoly law provide progressively more scope for the public interest. Computer trespass law favours data controllers. Contract law provides scope for gaming but is largely biased in favour of data controllers. Intellectual property law provides an explicit public interest exception to the rule. Unfair competition laws and anti-monopoly laws generally consider the impact on competition and consumers in the marketplace and weigh the public interest more broadly. The author believes that competition law is a more appropriate measure of private and public interest in cases where data ownership is unclear, and this is consistent with the trend of US courts' gradually limiting the application of the CFAA in data scraping cases.

Second, the public interests of privacy, personal information protection and data security can be incidental factors for consideration only under competitive effects in disputes between data controllers and scrapers. The case review shows that the conducts related to data protection are not per se harmful to competition, while their competition effects are different depending on the facts of the cases. Consumer-related or data security laws already provide sufficient legal remedies if data scrapers infringe on the privacy of users or consumers. Considering the protection of data-related rights and privacy separately in competitive disputes may lead to the risk of hollowing out the data protection law. Indeed, several empirical studies of the effects of the GDRP have shown that overprotection of personal data can lead to the creation of more concentrated market structures, preventing the realization of some data synergies and the losses of consumers (Aridor *et al.* 2023). If a data controller seeks relief on the grounds of data protection and privacy, it needs to demonstrate that its data control measures improve market transparency, increase inter-brand competition and do not merely disrupt the market in a broad sense. Third, for a data user to claim an exemption on the basis of obtaining user authorization or data portability, it needs to demonstrate that the use of the data increases consumer choice or facilitates consumer multihoming switching or that the use of the data generates innovations that promote economic efficiency and can be shared by consumers.

Finally, the public interest in data flow and free competition is also fundamental to freedom of expression. Data itself are carriers of expression, and being able to access and use data are naturally prerequisites for freedom of expression. However, the direct consideration of freedom of expression in data scraping disputes is a territorial issue. In reality, only the broad scope of unconstitutional review under US law provides the possibility to consider this issue in specific cases. US judges also avoid discussing complex constitutional issues in real-world cases. In *Sandvig v. Barr*, the court apparently avoided delving into issues relating to First Amendment protections. In short, the current data scraping dispute is not directly related to freedom of expression, and it is not feasible to consider this public interest in specific cases.

Based on the preceding reflections, it is suggested that a further public interest exception be created in cases where there is substantial injury so that courts can consider the reasonableness of the scraping conduct or arguments about pro-competitive effects in the context of competition law. The burden of such proof of public interest should be imposed on the scrapers, who could use it as an affirmative defence to justify the scraping conduct. The public interest should be specific, empirically based and clear rather than general. For this purpose, it is extremely important for courts to fully understand the economic incentives of the parties and their associated conduct as well as the factors affecting the market and the static and dynamic effects of the conduct. It is submitted that this analysis might require an in-depth understanding of the market dynamics that can be run within an effects or an efficiency analysis framework rather than any legal right or contractual authorization. The privacy or security issue, rather than a single public interest, should be included in the competitive analysis. The data controllers assume the burden of proving that there is substantial injury caused by scraping conduct. If a scraper does not cause substantial injury to the data holder, then the scraping conduct should be viewed as legal. For example, in the cases of search engines, aggregated websites and price comparison sites, data crawling simply makes it easier for Internet users to access information and ultimately to access the data

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

appropriator's websites and platforms to receive services or purchase goods through the relevant indexing or aggregation sites.

6 Conclusion

One of the great advantages of the Internet is that it provides easy access to data through powerful search facilities, ease of communication, and sophisticated analysis capabilities. Data scraping as a technical tool plays an important role in obtaining and using data. Like any other tool, data scraping can be used for good or bad. The diversity of objects and scenarios in which this technology is used makes it difficult to apply the law. This article has mapped the trajectory of relevant laws and jurisprudence around the legality of data scraping in the US, the EU and China, and it has been found that courts have made different decisions based on different public policy preferences. We believe that courts should not ignore the public nature of Internet data and the original intention of free sharing and innovation on the Internet when handling data scraping cases but should take the maintenance of data flow as the main consideration of interest.

References

Andrew P (2022) Unfair collection: Reclaiming control of publicly available personal information from data scrapers. *Michigan Law Review* **120**, 913.

Aridor G, Che Y-K and Salz T (2023) The effect of privacy regulation on the data industry: Empirical evidence from GDPR. *The RAND Journal of Economics* **54**, 695–730.

Brown RS (1948) Advertising and the public interest: Legal protection of trade symbols. *The Yale Law Journal* **57**, 1165–1206.

Dreyer AJ and Stockton J (2013) Internet 'data scraping': A primer for counseling clients. *New York Law Journal*. Available at https://www.law.com/newyorklawjournal/almID/1202610687621/Internet-'Data-Scraping'%3A-A-Primer-for-Counseling-Clients (accessed 15 July 2013).

Elkin-Koren N (2020) Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. *Big Data and Society* **7**, 2053951720932296.

Gold Z and Latonero M (2017) Robots welcome: Ethical and legal considerations for web crawling and scraping. *Washington Journal of Law, Technology and Arts* **13**, 275.

Goldfein S and Keyte J (2017) Big data, web 'scraping' and competition law: The debate continues. *New York Law Journal* **258**, 1–3.

Grover V, Chiang RHL, Liang T-P and Zhang D (2018) Creating strategic business value from big data analytics: A research framework. *Journal of Management Information Systems* **35**, 388–423.

Hirschey JK (2014) Symbiotic relationships: Pragmatic acceptance of data scraping. *Berkeley Technology Law Journal* **29**, 897.

Ingrams A (2019) Public values in the age of big data: A public information perspective. *Policy and Internet* **11**, 128–148.

Ioannis D (2019) Liability for data scraping prohibitions under the refusal to deal doctrine: An incremental step toward more robust Sherman Act enforcement. *University of Chicago Law Review* **86**, 1901.

ITR (2022) *2022 Imperva Bad Bot Report*. San Mateo, CA: Imperva, Inc.

Jamie WL (2018) Automation is not hacking: Why courts must reject attempts to use the CFAA as an anti competitive sword. *Boston University Journal of Science and Technology Law* **42**, 416.

Kerr OS (2016) Norms of computer trespass. *Columbia Law Review* **116**, 1143–1183.

Krämer J and Wohlfarth M (2018) Market power, regulatory convergence, and the role of data in digital markets. *Telecommunications Policy* **42**, 154–171.

Krotov V and Johnson L (2022) Big web data: Challenges related to data, technology, legality, and ethics. *Business Horizons* **66**, 481–491.

Lemley MA (2005) Property, intellectual property, and free riding. *Texas Law Review* **83**, 1031.

Luscombe A, Dick K and Walby K (2022) Algorithmic thinking in the public interest: Navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality and Quantity* **56**, 1023–1044.

Mancosu M and Vegetti F (2020) What you can scrape and what is right to scrape: A proposal for a tool to collect public Facebook data. *Social Media + Society* **6**, 2056305120940703.

Martens B and Zhao B (2021) Data access and regime competition: A case study of car data sharing in China. *Big Data and Society* **8**, 20539517211046374.

Mayernik MS (2017) Open data: Accountability and transparency. *Big Data and Society* **4**, 2053951717718853.

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press

McRory WC (2021) Let the bots be bots: Why the CFAA must be clarified to prevent the selective banning of data collection facilitating private social media information monopolization. *Brooklyn Journal of Corporate, Financial and Commercial Law* **16**, 279.

Najork M (2009) Web crawler architecture. *Encyclopedia of Database Systems*. New York: Springer.

Rahman RU and Tomar DS (2021) Threats of price scraping on e-commerce websites: Attack model and its detection using neural network. *Journal of Computer Virology and Hacking Techniques* **17**, 75–89.

Riley KC (2018) Data scraping as a cause of action: Limiting use of the CFAA and trespass in online copying cases. *Fordham Intellectual Property, Media and Entertainment Law Journal* **29**, 245.

Sellars A (2018) Twenty years of web scraping and the Computer Fraud and Abuse Act. *Boston University Journal of Science & Technology Law* **24**, 381–388.

Sobel B (2020) A new common law of web scraping. *Lewis & Clark Law Review* **25**, 147.

Tsang KF (2022) International application of CFAA: Scraping data or scraping law? *St. Louis University Law Journal* **66**, 1–27.

Udapure TV, Kale RD and Dharmik RC (2014) Study of web crawler and its different types. *IOSR Journal of Computer Engineering* **16**, 1–5.

Vezyridis P and Timmons S (2019) Resisting big data exploitations in public healthcare: Free riding or distributive justice? *Sociology of Health and Illness* **41**, 1585–1599.

Wierzel KL (2001) If you can't beat them, join them: data aggregators and financial institutions. *North Carolina Banking Institute* **5**, 457.

Xiao G (2020) Bad bots: Regulating the scraping of public personal information. *Harvard Journal of Law and Technology* **34**, 701.

Cite this article: Fei L (2024). A comparative study on public interest considerations in data scraping dispute. *International Journal of Law in Context* **20**, 568–580. https://doi.org/10.1017/S1744552324000156

https://doi.org/10.1017/S1744552324000156 Published online by Cambridge University Press